NCBI dbSNP bitfield encoding schema

Last two revisions:
version 4, Rev. May 2008
version 3, Rev. October 2007

12 byte structure

encoding	links	gene function	mapping	frequency	genotype	HapMap	phenotype	class	quality
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9

Low order byte   byte 1   byte 2   byte 1   byte 2   High order byte
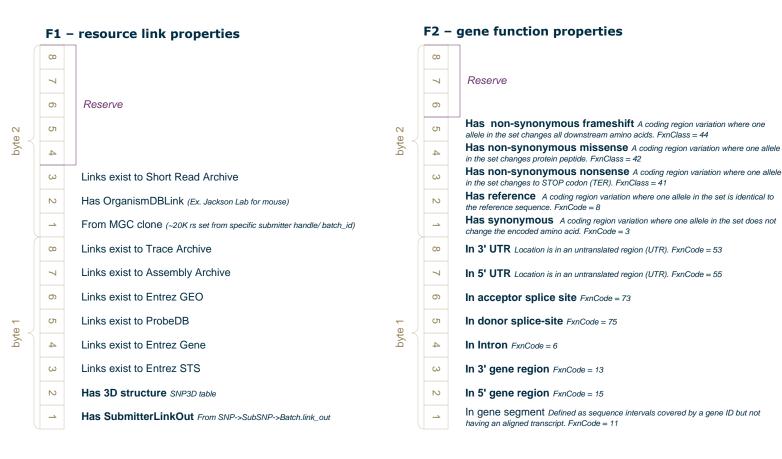
The dbSNP bitfield structure is a 12-byte object that defines variation properties in 9 areas.
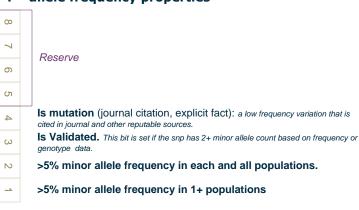- F1 and F2 are 2-byte structures
- F0 (lowest order byte) is the version of the encoding schema used for the data, see page 2 for definition.
- Bits labeled in **bold** are currently populated by dbSNP

F1 – resource link properties

byte 2

8	
7	*Reserve*
6	
5	
4	
3	Links exist to Short Read Archive
2	Has OrganismDBLink *(Ex. Jackson Lab for mouse)*
1	From MGC clone *(~20K rs set from specific submitter handle/ batch_id)*

byte 1

8	Links exist to Trace Archive
7	Links exist to Assembly Archive
6	Links exist to Entrez GEO
5	Links exist to ProbeDB
4	Links exist to Entrez Gene
3	Links exist to Entrez STS
2	**Has 3D structure** *SNP3D table*
1	**Has SubmitterLinkOut** *From SNP->SubSNP->Batch.link_out*

F2 – gene function properties

byte 2

8	
7	*Reserve*
6	
5	**Has non-synonymous frameshift** *A coding region variation where one allele in the set changes all downstream amino acids. FxnClass = 44*
4	**Has non-synonymous missense** *A coding region variation where one allele in the set changes protein peptide. FxnClass = 42*
3	**Has non-synonymous nonsense** *A coding region variation where one allele in the set changes to STOP codon (TER). FxnClass = 41*
2	**Has reference** *A coding region variation where one allele in the set is identical to the reference sequence. FxnCode = 8*
1	**Has synonymous** *A coding region variation where one allele in the set does not change the encoded amino acid. FxnCode = 3*

byte 1

8	**In 3' UTR** *Location is in an untranslated region (UTR). FxnCode = 53*
7	**In 5' UTR** *Location is in an untranslated region (UTR). FxnCode = 55*
6	**In acceptor splice site** *FxnCode = 73*
5	**In donor splice-site** *FxnCode = 75*
4	**In Intron** *FxnCode = 6*
3	**In 3' gene region** *FxnCode = 13*
2	**In 5' gene region** *FxnCode = 15*
1	In gene segment *Defined as sequence intervals covered by a gene ID but not having an aligned transcript. FxnCode = 11*

F3 – mapping properties

8	
7	*Reserve*
6	
5	**Has other snp** *with exactly the same set of mapped positions on NCBI refernce assembly.*
4	**Has Assembly conflict.** *This is for weight 1 and 2 snp that maps to different chromosomes on different assemblies.*
3	**Is Assembly specific.** *This bit is 1 if the snp only maps to one assembly.*
2	**Weight (2 bits). Weight on NCBI reference assembly.** *map weight (2-bit: 00 = unmapped, 01 = 1, 10 = 2, 11 = 3 or more).*
1	

F4 – allele frequency properties

8	
7	*Reserve*
6	
5	
4	**Is mutation** (journal citation, explicit fact): *a low frequency variation that is cited in journal and other reputable sources.*
3	**Is Validated.** *This bit is set if the snp has 2+ minor allele count based on frequency or genotype data.*
2	**>5% minor allele frequency in each and all populations.**
1	**>5% minor allele frequency in 1+ populations**

NCBI dbSNP bitfield encoding schema
version 3, Rev. October 2007

F5 – genotype properties

8	
7	*Reserve*
6	
5	
4	
3	**Marker is on high density genotyping kit** *(50K density or greater). The snp may have phenotype associations present in dbGaP.*
2	In Haplotype tagging set
1	**Genotypes available.** *The snp has individual genotype (in SubInd table).*

F6 – HapMap properties

8	
7	*Reserve*
6	
5	
4	
3	Phase 3 genotyped: filtered, non-redundant.
2	**Phase 2 genotyped:** filtered, non-redundant.
1	Phase 1 genotyped: filtered, non-redundant.

F7 – phenotype properties

8	Has MeSH is linked to a disease.
7	**Variation is interrogated in a clinical diagnostic assay**
6	Has transcription factor
5	**Submitted from a locus-specific database.**
4	Has significant association in dbGaP study
3	Has LOD score
2	Has SnpRIF
1	**Has OMIM/OMIA**

F8 – variation class

8	
7	*Reserve*
6	
5	
4	0001 = single base polymorphism SNP
3	0010 = dips (deletion/insertion) 0011 = HETEROZYGOUS 0100 = Microsatellite
2	0101 = Named variation, e.g. (Alu) 0110 = NOVARIATION 0111 = mixed class
1	1000 = multi-base polymorphism

F9 – quality check

8	
7	*Reserve*
6	
5	**Contig allele not present in SNP allele list**. *The reference sequence allele at the mapped position is not present in the SNP allele list, adjusted for orientation.*
4	**Is Withdrawn by submitter** *If one member ss is withdrawn by submitter, then this bit is set. If all member ss' are withdrawn, then the rs is deleted to SNPHistory.*
3	**Rs cluster has non-overlapping allele sets**. *True when rs set has more than 2 alleles from different submissions and these sets share no alleles in common.*
2	Is a strain-specific fixed difference
1	**Has Genotype Conflict** *Same (rs, ind), different genotype. N/N is not included.*

F0 – Version encoding

8	
7	*Reserve*
6	
5	
4	
3	
2	**Bitmap schema version.** *Versions increment as integer value (NOTE: version 1.2 is encoded as version 1)*
1	